

## Choice of Scores in Trend Tests for Case-Control Studies of Candidate-Gene Associations

GANG ZHENG<sup>1</sup>, BORIS FREIDLIN<sup>2</sup>, ZHAOHAI LI<sup>3</sup>, and JOSEPH L. GASTWIRTH<sup>3</sup>

<sup>1</sup> Office of Biostatistics Research, National Heart, Lung and Blood Institute, 6701 Rockledge Drive, MSC 7938, Bethesda, MD 20892, U.S.A.

<sup>2</sup> Biometric Research Branch, National Cancer Institute, Executive Plaza North, MSC 7434, Bethesda, MD 20892, U.S.A.

<sup>3</sup> Biostatistics Branch, National Cancer Institute and Department of Statistics, George Washington University, Washington, DC 20052, U.S.A.

### *Abstract*

When applying the Cochran-Armitage (CA) trend test for an association between a candidate allele and a disease in a case-control study, a set of scores must be assigned to the genotypes. SASIENI (1997, *Biometrics* **53**, 1253–1261) suggested scores for the recessive, additive, and dominant models but did not examine their statistical properties. Using the criteria of minimizing the required sample size of the CA trend test to achieve prespecified type I and type II errors, we show that the scores given by SASIENI (1997) are optimal for the recessive and dominant models and locally optimal for the additive one. Moreover, the additive scores are shown to be locally optimal for the multiplicative model. The tests are applied to a real dataset.

**Key words:** Association; Case-control; Cochran-Armitage trend test; Optimal score.

### 1. Introduction

To test for an association between a candidate allele and disease, SASIENI (1997) showed that genotype-based Cochran-Armitage (CA) trend tests (COCHRAN, 1954; ARMITAGE, 1955) are preferable to allele-based tests as they are valid whether or not Hardy-Weinberg equilibrium (HWE) holds. To apply the CA trend test, scores are assigned to each of these genotypes, where the choice of scores depends on the underlying genetic model (SASIENI, 1997). Properties of tests using one set of scores have been studied by SLAGER and SCHAID (2001) and FREIDLIN et al. (2002). The need to specify the scores when using CA trend tests is a major concern, when there is uncertainty about the model underlying the data (GRAUBARD and KORN, 1987; PODGOR et al., 1996; NEUHAUSER and HOTHORN, 1999).

For a given genetic model, we derive an optimal score, which minimizes the required sample size for the CA trend test of size  $\alpha$  to achieve power

\* Corresponding author: zhengg@nhlbi.nih.gov

$100(1 - \beta)\%$ . The scores for the recessive and dominant models given by SASIENI (1997) are shown to be optimal in this sense. For the additive model, however, the score given by SASIENI (1997) is only locally optimal. We reparameterize the usual family of genetic models by introducing a parameter determined by the underlying genetic model. The optimal scores are functions of unknown parameters. So they cannot be directly used in practice. Hence, we examine the performance of a locally optimal score relative to the optimal one for a specific non-local alternative. This locally optimal score does not depend on unknown parameters. Simulations for moderate sample sizes confirm that the additive scores are nearly as powerful as the optimal scores for the multiplicative model.

The paper is organized as follows. The CA trend test and sample size are reviewed in Section 2. The reparameterization is introduced in Section 3. Optimal scores are derived in Section 4. In Section 5, we compare the sample sizes based on the optimal scores and the locally optimal scores. Simulation results are also present in Section 5. A real dataset is analyzed in Section 6. Section 7 is a brief discussion.

## 2. Review of Cochran-Armitage Trend Tests and Sample Size Calculations

The data available from a case-control study for candidate-gene association are given in Table 1. Let  $A$  be a high risk candidate allele and  $a$  be a lower risk allele. In Table 1, we assume that  $(r_0, r_1, r_2)$  and  $(s_0, s_1, s_2)$  follow trinomial distributions with probabilities for the genotypes  $aa$ ,  $aA$  and  $AA$  equal to  $p_0, p_1, p_2$  and  $q_0, q_1, q_2$ , respectively. Let  $R = \sum_i r_i$  and  $S = \sum_i s_i$  be the sizes of two independent random samples of cases and controls, respectively. The total sample size is  $N = R + S$ . The null hypothesis of no association is  $H_0 : p_i = q_i$  for  $i = 0, 1, 2$ .

To test  $H_0$  using a CA trend test, a score  $x = (x_0, x_1, x_2)$  is assigned to the genotypes  $(aa, aA, AA)$  such that  $0 \leq x_0 \leq x_1 \leq x_2$  and  $x_2 > x_0$ . SASIENI (1997) assigned  $x = (0, 0, 1)$  to the recessive model,  $x = (0, 1, 2)$  to the additive model, and  $x = (0, 1, 1)$  to the dominant model. The intuition underlying the scores is the following: for the recessive (dominant) model, the relative risks of the genotypes  $aa$  ( $AA$ ) and  $aA$  are the same, so the same score is assigned to  $aa$  and  $aA$  (for the recessive model) or  $aA$  and  $AA$  (for the dominant model). For the additive model,

Table 1  
Genotype distribution

	$aa$	$aA$	$AA$	Total
Cases	$r_0$	$r_1$	$r_2$	$R$
Controls	$s_0$	$s_1$	$s_2$	$S$
Total	$n_0$	$n_1$	$n_2$	$N$

the effect of  $aA$  should be the average of the effect of  $aa$  and  $AA$ , which is satisfied when the scores equal the number of  $A$  alleles in the genotype.

Given a score  $x$ , the standardized CA trend test is  $Z = U / \{\widehat{\text{var}}_{H_0}(U)\}^{1/2}$  where  $U = \sum_{i=0}^2 x_i(Sr_i - Rs_i)/N$  and

$$\widehat{\text{var}}_{H_0}(U) = \frac{RS}{N^3} \left[ N \sum_{i=0}^2 x_i^2 n_i - \left( \sum_{i=0}^2 x_i n_i \right)^2 \right]. \quad (1)$$

Under the null hypothesis  $H_0$ ,  $Z$  has an asymptotic standard normal distribution. In Appendix A, we show that, under the alternative hypothesis  $H_a$ ,  $E_{H_a}(U) > 0$ . Thus, we consider one-sided CA trend tests, and reject the null hypothesis of no association when  $Z > z_{1-\alpha}$ , where  $z_p$  is the  $p$ th percentile of a standard normal distribution and  $\alpha$  is the significance level. The CA trend test is a function of the scores but it is invariant to a linear transformation of them (TARONE and GART, 1980). Hence, CA trend tests based on scores  $x = (x_0, x_1, x_2)$  and  $x = (0, \eta, 1)$ , where  $\eta = (x_1 - x_0)/(x_2 - x_0)$  are the same.

From FREIDLIN et al. (2002), for testing  $H_0 : p_i = q_i$  for  $i = 0, 1, 2$ , at significance level  $\alpha$  to achieve power equal to  $100(1 - \beta)\%$ , the required sample size for  $Z$  with the set of scores  $x = (x_0, x_1, x_2)$  is approximately

$$N = \left[ z_{1-\alpha} \left\{ 1 + \left( \frac{\tilde{\sigma}_a}{\mu_a} \right)^2 \right\}^{1/2} + z_{1-\beta} \left( \frac{\sigma_a}{\mu_a} \right) \right]^2, \quad (2)$$

where  $\mu_a = (RS/N^2) \sum_{i=0}^2 x_i(p_i - q_i)$ ,

$$\sigma_a^2 = \frac{RS^2}{N^3} \left[ \sum_{i=0}^2 x_i^2 p_i - \left( \sum_{i=0}^2 x_i p_i \right)^2 \right] + \frac{R^2 S}{N^3} \left[ \sum_{i=0}^2 x_i^2 q_i - \left( \sum_{i=0}^2 x_i q_i \right)^2 \right], \quad (3)$$

$$\tilde{\sigma}_a^2 = \frac{R^2 S}{N^3} \left[ \sum_{i=0}^2 x_i^2 p_i - \left( \sum_{i=0}^2 x_i p_i \right)^2 \right] + \frac{RS^2}{N^3} \left[ \sum_{i=0}^2 x_i^2 q_i - \left( \sum_{i=0}^2 x_i q_i \right)^2 \right]. \quad (4)$$

### 3. A Family of Genetic Models

Denote the penetrances for the genotypes  $aa$ ,  $aA$  and  $AA$  by  $f_0$ ,  $f_1$  and  $f_2$ , respectively, where  $f_i = \Pr(\text{case} \mid i \text{ A alleles})$ . As  $A$  is the candidate high risk allele, we assume  $f_2 \geq f_1 \geq f_0 > 0$ . Let  $h_i = 1 - f_i = \Pr(\text{control} \mid i \text{ A alleles})$ , for  $i = 0, 1, 2$ ; note that  $h_i$  is the probability of being non-diseased given  $i$   $A$  alleles. Let  $K$  denote the disease prevalence and  $\gamma_i = f_i/f_0$  and  $\delta_i = h_i/h_0$  be genotype relative risks,  $i = 1, 2$ . Define the population genotype probabilities as  $g_0 = \Pr(aa)$ ,

$g_1 = \Pr(aA)$  and  $g_2 = \Pr(AA)$ . With this notation,  $p_i$  and  $q_i$  can be expressed as

$$p_i = \frac{f_i g_i}{K} = \frac{\gamma_i g_i}{\sum_i \gamma_i g_i} \quad \text{and} \quad q_i = \frac{(1 - f_i) g_i}{1 - K} = \frac{\delta_i g_i}{\sum_i \delta_i g_i}, \quad (5)$$

where  $\gamma_0 = \delta_0 = 1$ . The null hypothesis of no association can also be stated as  $H_0 : \gamma_1 = \gamma_2 = 1$ , i.e.,  $H_0 : \delta_1 = \delta_2 = 1$ .

The three common genetic models are specified by  $\gamma_1 = 1$  for the recessive model;  $\gamma_1 = \gamma_2$  for the dominant model, and  $2\gamma_1 = 1 + \gamma_2$  for the additive model. Specific models with genotype relative risks  $\gamma_1$  and  $\gamma_2$  form a class denoted by  $\Omega_\gamma = \{(\gamma_1, \gamma_2) : 1 \leq \gamma_1 \leq \gamma_2\}$ . The null hypothesis corresponds to  $H_0 : (\gamma_1, \gamma_2) = (1, 1) \in \Omega_\gamma$  and the possible alternative hypotheses are  $H_a : (\gamma_1, \gamma_2) \in \Omega_\gamma - (1, 1)$ . In practice, we do not know the values of  $\gamma_1$  and  $\gamma_2$ . For recessive, dominant and additive models,  $\gamma_1$  and  $\gamma_2$  follow a linear relation, so they can be specified by a ray from the null value  $(1, 1)$  (Figure 1). In Figure 1,  $x$  ( $y$ ) axis corresponds to  $\gamma_1$  ( $\gamma_2$ ). From Figure 1, under the alternative hypothesis, possible recessive (REC), additive (ADD), and dominant (DOM) models correspond to points on the rays OA, OC, and OD, respectively.

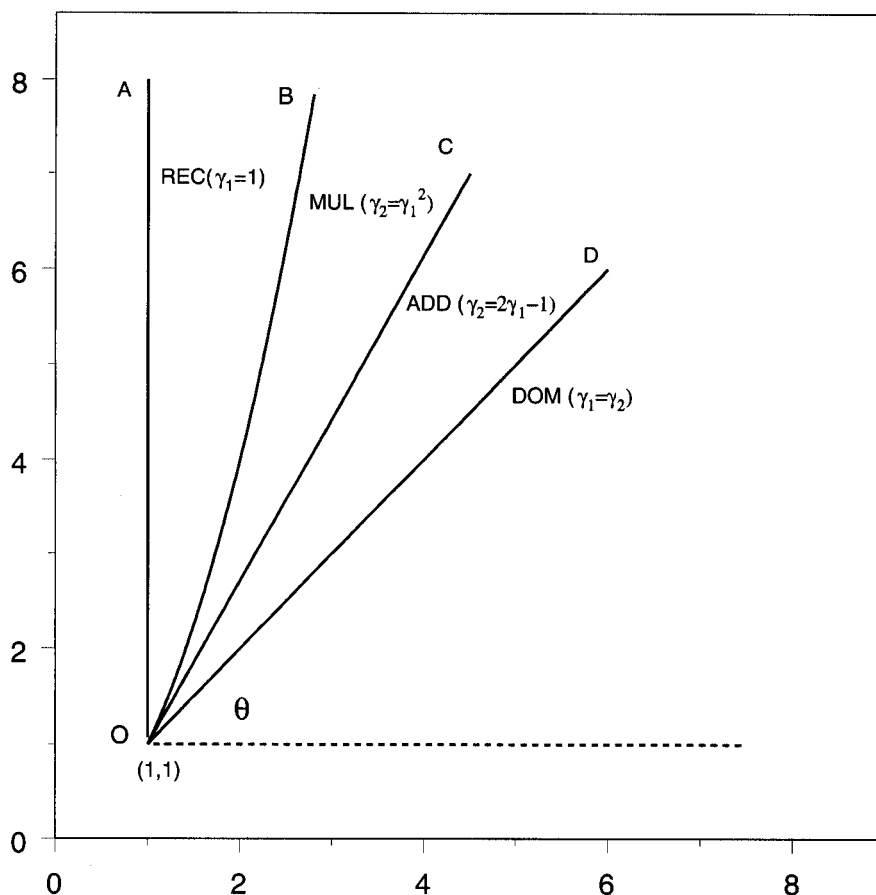


Fig. 1. The family of genetic models and the reparameterization

Note that a genetic model is a function of two relative risks  $\gamma_1$  and  $\gamma_2$ . For each genetic model, there is a corresponding optimal test statistic. We introduce a reparameterization, which simplifies the calculation of the optimal test statistic for any genetic model.

Let  $P = (\gamma_1, \gamma_2)$  be the true relative risks in  $\Omega_\gamma$  and  $(\gamma_1, \gamma_2) \neq (1, 1)$ . Let the distance between  $O = (1, 1)$  and  $P$  be  $r > 0$  and  $\theta$  be the angle between  $OP$  and the horizontal line (the dashed line in Figure 1). Then  $\gamma_1 = 1 + r \cos \theta$  and  $\gamma_2 = 1 + r \sin \theta$ , where  $r^2 = (\gamma_1 - 1)^2 + (\gamma_2 - 1)^2$  and  $\theta \in [\pi/4, \pi/2]$  (Figure 1). Thus we can express the null and alternative hypotheses as  $H_0 : r = 0$  and  $H_a : r > 0$ , respectively, where  $\theta$  is a parameter which is determined by the underlying genetic model. Given any  $(\gamma_1, \gamma_2) \in \Omega_\gamma$  and  $(\gamma_1, \gamma_2) \neq (1, 1)$ ,  $\theta = \cot^{-1}[(\gamma_1 - 1)/(\gamma_2 - 1)]$ . For example, we have  $\theta = \pi/2$  for the recessive model,  $\theta = \cot^{-1}(1/2)$  ( $= 63.5^\circ$ ) for the additive model, and  $\theta = \pi/4$  for the dominant model. Hence, we only need to derive an optimal test statistic assuming  $\theta$  is known. For a specific genetic model, we replace  $\theta$  with the corresponding known value to obtain the optimal test statistic for that genetic model. Further, when the underlying genetic model is unknown, robust inference for candidate-gene association between genetic marker and trait can be focused on this family of genetic models where  $\theta$  can be treated as an unknown parameter (e.g., WHITTEMORE and TU, 1998; GASTWIRTH and FREIDLIN, 2000; SHIH and WHITTEMORE, 2001; FREIDLIN et al., 2002). More discussion of this reparameterization is given in Appendix B.

#### 4. Optimal Scores for Cochran-Armitage Trend Tests

For the genetic model defined by a specific  $\theta \in [\pi/4, \pi/2]$ , a score  $x = (0, \eta, 1)$ ,  $\eta \in [0, 1]$  is optimal if the sample size  $N$  for a test based on it minimizes the required sample size for any CA trend test ( $Z$ ) achieving the same power. Since  $N$  given by (2) is a continuous function of  $\eta \in [0, 1]$ , there exists an  $\eta \in [0, 1]$  at which  $N$  is minimized. By a numerical search over  $\eta = 0(0.001)1$ , one obtains the optimal score, defined by  $\eta$ , that minimizes the sample size for the assumed genetic model.

When  $R = S$ , analytical results can be found. From (3) and (4), if  $R = S$ ,  $\sigma_a^2 = \tilde{\sigma}_a^2$ . If  $\mu_a^2 \ll \sigma_a^2$ , then we can apply a Taylor expansion to  $\{1 + (\mu_a/\sigma_a)^2\}^{1/2}$  in (2), yielding  $N \approx (z_{1-\alpha} + z_{1-\beta})^2 (\sigma_a/\mu_a)^2 + z_{1-\alpha}(z_{1-\alpha} + z_{1-\beta})$ , where the error term of the approximation is  $O((\mu_a/\sigma_a)^4)$ . For most genetic applications, FREIDLIN et al. (2002) showed that (2) and the approximation yield the exact same sample sizes for any scores. The benefit of using the approximation is that, if  $R = S$ , we only need to minimize  $\sigma_a^2/\mu_a^2$  to find the optimal score.

The proof of the following result is given in Appendix C.

**Theorem 1:** Under  $H_0$ , for any given  $\theta \in [\pi/4, \pi/2]$  and score  $x = (0, \eta, 1)$ ,  $\sigma_a^2/\mu_a^2$  is minimized for any  $R$  and  $S$  if and only if  $\eta = \eta^*$ , where

$$\eta^* = \frac{RK(1 - f_2) + Sf_2(1 - K)}{RK(1 - f_1) + Sf_1(1 - K)} \cot \theta. \quad (6)$$

Moreover, if  $R = S$ ,  $N$  is minimized when  $\eta = \eta^*$ , a strictly convex function of  $\eta$ .

**Corollary 1:** Under the same conditions of Theorem 1,  $\tilde{\sigma}_a^2/\mu_a^2$  is minimized for any  $R$  and  $S$  if and only if  $\eta = \eta^{**}$ , where

$$\eta^{**} = \frac{SK(1 - f_2) + Rf_2(1 - K)}{SK(1 - f_1) + Rf_1(1 - K)} \cot \theta. \quad (7)$$

The proof of Corollary 1 is similar to that of Theorem 1.

Since  $K = \sum_i f_i g_i$  and  $1 - K = \sum_i (1 - f_i) g_i$ , from (5),  $f_i/(1 - f_i) = Kp_i/[(1 - K)q_i]$ , i.e.,  $f_i = Kp_i/[q_i + K(p_i - q_i)]$ . Hence (6) can also be expressed as

$$\eta^* = \left( \frac{Rq_2 + Sp_2}{Rq_1 + Sp_1} \right) \left[ \frac{q_1 + K(p_1 - q_1)}{q_2 + K(p_2 - q_2)} \right] \cot \theta. \quad (8)$$

When  $R = S$ , Theorem 1 gives an optimal score  $x^* = (0, \eta^*, 1)$  that minimizes the required sample size  $N$  of the CA trend test achieving a pre-set size and power of the test for the genetic model specified by  $\theta$ . The optimal scores, however, depend on unknown parameters  $K$  and  $p_i, q_i$ , so they cannot be calculated. The scores of SASIENI (1997) for the recessive, additive (multiplicative), and dominant models can be written as  $x = (0, \cot \theta, 1)$ , where  $\theta = \pi/2, \cot^{-1}(1/2)$ , and  $\pi/4$ , respectively. For any given  $\theta \in [\pi/4, \pi/2]$ , we are interested in when the scores  $x = (0, \cot \theta, 1)$  can be used. The following result gives the condition (the proof is given in Appendix D).

**Theorem 2:** Under  $H_a$ ,  $\theta \in [\pi/4, \pi/2]$  is given. Consider the score  $x = (0, \eta, 1)$ . For any  $R$  and  $S$ ,

- (i) If  $\theta = \pi/4$  or  $\theta = \pi/2$ ,  $\sigma_a^2/\mu_a^2$  is minimized if and only if  $\eta = \cot \theta$ .
- (ii) If  $\theta \in (\pi/4, \pi/2)$ , say,  $\theta = \cot^{-1}(1/2)$  (the additive model), and  $\eta = \cot \theta$  is used, then  $\sigma_a^2/\mu_a^2$  is minimized if and only if  $R/S = (1 - K)/K$ .

From Theorem 2, when  $R = S$ , the scores  $x = (0, 0, 1)$  and  $x = (0, 1, 1)$  of SASIENI (1997) for the recessive and dominant models are optimal. For the additive model, when  $R = S$ , the score  $x = (0, 1/2, 1)$  used by SASIENI (1997) is not necessarily optimal. In the following, we consider the recessive and dominant models for any  $R$  and  $S$ . Note that  $N$  given by (2) depends on  $\sigma_a/\mu_a > 0$  and  $\tilde{\sigma}_a/\mu_a > 0$ . The score yielding a test that minimizes  $\sigma_a^2/\mu_a^2$  ( $\tilde{\sigma}_a^2/\mu_a^2$ ) is given by  $\eta^*$  ( $\eta^{**}$ ). When  $\eta^* = \eta^{**}$ , that score minimizes  $N$  and is optimal. From (6) and (7), when  $\theta = \pi/2$  (recessive model), both  $\tilde{\sigma}_a^2/\mu_a^2$  and  $\sigma_a^2/\mu_a^2$  are minimized when  $x = (0, 0, 1)$ , i.e.,  $\eta^* = \eta^{**} = 0$ . On the other hand, if  $\theta = \pi/4$  (dominant model), then  $f_1 = f_2$ , so  $\eta^* = \eta^{**} = 1$ . Hence we obtain:

**Corollary 2:** Under  $H_a$ , the scores  $x = (0, 0, 1)$  and  $x = (0, 1, 1)$  are optimal for the recessive and dominant models, respectively, for all values of  $R$  and  $S$ .

For  $\theta \in (\pi/4, \pi/2)$  and any  $R$  and  $S$ , we consider a local property of the score  $x = (0, \cot \theta, 1)$ . Under  $H_0$ ,  $p_i = q_i$  for  $i = 0, 1, 2$ . Hence, defining a local alternative hypothesis as  $p_i = q_i + c_i N^{-1/2} = q_i + O(N^{-1/2})$  for some constant  $c_i$  and  $i = 0, 1, 2$ , we have:

**Theorem 3:** Suppose  $p_i = q_i + O(N^{-1/2})$  and  $\theta \in (\pi/4, \pi/2)$  is given. Let  $x^* = (0, \eta^*, 1)$  and  $x^{**} = (0, \eta^{**}, 1)$  be the scores minimizing  $\sigma_a^2/\mu_q^2$  and  $\tilde{\sigma}_a^2/\mu_a^2$ , respectively. Then  $\eta^* = \cot \theta + O(N^{-1/2})$  and  $\eta^{**} = \cot \theta + O(N^{-1/2})$  and the score  $x = (0, \cot \theta, 1)$  is locally optimal for  $\theta \in (\pi/4, \pi/2)$  for all values of  $R$  and  $S$ .

We describe the optimal scores (6) under  $R = S$  and HWE for the additive model ( $\cot \theta = 0.5$ ) and other genetic models defined by  $\cot \theta$ . We choose  $r = 0.5, 1.0, 2.0, 5.0$  for alternative hypotheses, and  $\cot \theta = 0.05, 0.20, 0.33, 0.50, 0.75, 0.95$ . Note that  $\cot \theta = 0.05$  and  $0.95$  are close to the recessive and dominant models, respectively, when  $r$  is close to 0. The values of  $(\gamma_1, \gamma_2)$  calculated from  $\gamma_1 = 1 + r \cos \theta$  and  $\gamma_2 = 1 + r \sin \theta$  are given in Table 2.

Table 2 shows that as  $r$  increases the models specified by  $\cot \theta = 0.05$  and  $0.95$  move further from the fully recessive ( $\gamma_1 = 1$ ) and dominant ( $\gamma_1 = \gamma_2$ ) ones, respectively. However, for moderate values of  $r$ , the genotype relative risks remain near to their values under the dominant and recessive models. Thus, locally optimal tests should have reasonable power for models in a fairly large neighborhood of the null.

Optimal scores for various models are presented in Table 3. The results in Table 3 show that the optimal scores for a specific genetic model hardly change as  $K$  and  $p$  vary. For local alternatives, e.g.,  $r = 0.5$ ,  $\eta^* \approx \cot \theta$ . Overall, the optimal score  $\eta^*$  is an increasing function of  $r$  when  $K$  and  $p$  are fixed. When  $r$  is fixed,  $\eta^*$  is a decreasing function of  $K$  or  $p$ . When  $r$  is at least 2.0, the difference between  $\eta^*$  and  $\cot \theta$  is noticeable for the additive ( $\cot \theta = 0.50$ ) and near recessive models ( $\cot \theta = 0.05$ ), while the value of  $\eta^*$  for near dominant models ( $\cot \theta = 0.95$ ) remains close to the locally optimal  $\eta = \cot \theta$ .

Table 2

Genotype relative risks  $(\gamma_1, \gamma_2)$  for genetic models specified by  $r$  and  $\theta$

$r$	True model $\cot \theta$					
	0.05 ( $\gamma_1, \gamma_2$ )	0.20 ( $\gamma_1, \gamma_2$ )	0.33 ( $\gamma_1, \gamma_2$ )	0.50 ( $\gamma_1, \gamma_2$ )	0.75 ( $\gamma_1, \gamma_2$ )	0.95 ( $\gamma_1, \gamma_2$ )
0.5	(1.025, 1.5)	(1.098, 1.49)	(1.157, 1.475)	(1.224, 1.447)	(1.3, 1.4)	(1.345, 1.363)
1.0	(1.05, 2.0)	(1.196, 1.98)	(1.313, 1.948)	(1.447, 1.894)	(1.6, 1.8)	(1.689, 1.725)
2.0	(1.10, 3.0)	(1.392, 2.96)	(1.626, 2.897)	(1.894, 2.788)	(2.2, 2.6)	(2.378, 2.451)
5.0	(1.25, 6.0)	(1.980, 5.90)	(2.565, 5.742)	(3.235, 5.470)	(4.0, 5.0)	(4.490, 4.670)

Table 3

Optimal scores  $\eta^*$  for the given model when  $R = S$  and HWE holds

$r$	$K$	$p$	True model $\cot \theta$					
			0.05	0.20	0.33	0.50	0.75	0.95
0.5	0.01	0.01	0.062	0.237	0.379	0.551	0.782	0.958
		0.50	0.061	0.234	0.375	0.546	0.779	0.957
	0.10	0.01	0.060	0.233	0.374	0.546	0.779	0.957
		0.50	0.060	0.231	0.370	0.542	0.776	0.956
1.0	0.01	0.01	0.073	0.271	0.420	0.590	0.807	0.963
		0.50	0.070	0.261	0.407	0.576	0.798	0.961
	0.10	0.01	0.071	0.264	0.412	0.583	0.802	0.962
		0.50	0.068	0.255	0.399	0.569	0.793	0.960
2.0	0.01	0.01	0.095	0.330	0.488	0.652	0.842	0.970
		0.50	0.085	0.301	0.452	0.617	0.821	0.965
	0.10	0.01	0.090	0.318	0.475	0.641	0.836	0.969
		0.50	0.081	0.290	0.439	0.605	0.814	0.964
5.0	0.01	0.01	0.154	0.460	0.620	0.760	0.898	0.982
		0.50	0.115	0.365	0.517	0.671	0.849	0.972
	0.10	0.01	0.145	0.441	0.602	0.746	0.891	0.980
		0.50	0.106	0.346	0.497	0.654	0.840	0.970

## 5. Comparison of the Required Sample Size and Power of the Locally Optimal and Optimal Scores

As the sample size formula (2) and its approximation are based on large sample theory, the loss of power incurred by using the locally optimal score  $(0, \cot \theta, 1)$  instead of the optimal one should be examined. First, we compare sample sizes based on different scores when  $\cot \theta = 0.05, 0.20, 0.33, 0.50, 0.75, 0.95$  and  $r = 2.0, 5.0$  when  $R = S$  and HWE holds. Table 4 reports the results. When  $\cot \theta$  is relative large, say,  $\cot \theta > 0.20$ , the sample sizes based on the locally optimal scores  $x = (0, \cot \theta, 1)$  are very close to the sample sizes based on the optimal scores  $x^* = (0, \eta^*, 1)$ . When  $\cot \theta = 0.05$  and  $p = 0.01$ , the sample size based on the optimal score is somewhat smaller than the sample size based on the locally optimal score. Overall, for a high risk allele of moderate frequency, say,  $p \geq 0.1$ , the sample sizes based on the optimal and locally optimal scores are quite close. As optimal scores cannot be calculated in most applications, Table 4 suggests that, given the genetic model, the locally optimal score can be used when  $p \geq 0.1$ .

When  $R \neq S$ , neither  $x^* = (0, \eta^*, 1)$  nor  $x^{**} = (0, \eta^{**}, 1)$  are the optimal score for  $N$ . In this case, we compare the locally optimal score  $x = (0, \cot \theta, 1)$  and the optimal score  $x = (0, \eta_{\text{opt}}, 1)$  obtained by numerical search over  $\eta = 0(0.0001)1$  for a given model, the same parameters as Table 4 are used and  $R/N = 0.35$ . The results are similar to those in Table 4, so they are not reported here. When  $R = S$

and  $p \geq 0.1$ , the sample sizes based on locally optimal scores  $x = (0, \cot \theta, 1)$  are quite close to those based on optimal scores.

We conducted a simulation comparing the CA tests with the optimal score and the locally optimal score when  $R = S$  for the additive and multiplicative models. We assume the values of  $\gamma_1$ , the prevalence  $K$  and allele frequency  $p$  are known, and calculate marginal probabilities for genotypes  $g_i$  under HWE and  $\gamma_2 = 2\gamma_1 - 1$  (the additive model) and  $\gamma_2 = \gamma_1^2$  (the multiplicative model). The values of  $f_i$  are then calculated by  $f_0 = K / \sum_i \gamma_i g_i$  and  $f_i = \gamma_i f_0$ . We generate two independent trinomial distributions of sizes  $R$  and  $S$  ( $R = S = N/2$ ) with probabilities  $p_i$  and  $q_i$ ,  $i = 0, 1, 2$ , given by (5). For the additive model, the CA trend tests are calculated using the optimal score  $\eta^*$  given by (8), where  $\cot \theta = 0.5$  and the locally optimal score  $\eta = 0.5$ . For the multiplicative model, we calculate CA trend tests based on the optimal score  $\eta^*$  given by (8), where  $\cot \theta = (1 + \gamma_1)^{-1}$ , the locally optimal score  $\eta_e = \cot \theta = (1 + \gamma_1)^{-1}$  and the additive score  $\eta_a = 0.5$ .

Table 4

Sample sizes to achieve 80% power for one-sided CA trend test for the size  $\alpha = 0.05$  when  $R = S$  and HWE holds

$r$	$K$	$p$	$\cot \theta = 0.05$		$\cot \theta = 0.20$		$\cot \theta = 0.33$	
			$\eta = 0.05$	$\eta^*$	$\eta = 0.20$	$\eta^*$	$\eta = 0.33$	$\eta^*$
2.0	0.01	0.01	68,902	63,019	9,307	9,072	4,118	4,090
		0.50	100	100	140	139	175	171
	0.10	0.01	54,594	50,566	7,524	7,360	3,319	3,300
		0.50	81	81	114	113	144	141
5.0	0.01	0.01	17,786	13,556	1,998	1,859	919	903
		0.50	39	39	60	58	77	73
	0.10	0.01	13,757	10,796	1,584	1,484	726	714
		0.50	31	31	48	47	63	61

$r$	$K$	$p$	$\cot \theta = 0.50$		$\cot \theta = 0.75$		$\cot \theta = 0.95$	
			$\eta = 0.50$	$\eta^*$	$\eta = 0.75$	$\eta^*$	$\eta = 0.95$	$\eta^*$
2.0	0.01	0.01	2,255	2,252	1,397	1,397	1,123	1,123
		0.50	210	204	226	224	222	222
	0.10	0.01	1,807	1,805	1,112	1,112	891	891
		0.50	174	170	190	189	188	188
5.0	0.01	0.01	537	535	355	355	290	290
		0.50	94	88	99	97	97	96
	0.10	0.01	421	420	276	276	225	225
		0.50	78	74	85	83	83	83

$\eta^*$  is the optimal score for the given model (Table 3).

$\eta$  is the locally optimal score for the given model.

Table 5

Empirical power of one-sided CA trend tests with the optimal and locally optimal scores for the size  $\alpha = 0.05$  when  $R = S$  and HWE holds

$\gamma_1$	$K$	$p$	$N$	Additive		Multiplicative		
				$\eta^*$	$\eta$	$\eta^*$	$\eta_e$	$\eta_a$
1.0	0.01	0.05	200	0.051	0.051	0.052	0.052	0.052
		0.10	200	0.049	0.049	0.052	0.052	0.052
		0.30	200	0.053	0.053	0.051	0.051	0.051
2.0	0.01	0.05	200	0.529	0.527	0.555	0.550	0.554
		0.10	200	0.728	0.723	0.777	0.763	0.776
		0.30	200	0.886	0.879	0.958	0.948	0.958
	0.10	0.05	200	0.611	0.611	0.639	0.632	0.638
		0.10	200	0.809	0.806	0.847	0.842	0.847
		0.30	200	0.936	0.932	0.981	0.978	0.981
3.0	0.01	0.05	100	0.661	0.656	0.703	0.683	0.700
		0.10	100	0.832	0.828	0.891	0.857	0.893
		0.30	100	0.916	0.902	0.986	0.976	0.985
	0.10	0.05	100	0.749	0.747	0.795	0.792	0.795
		0.10	100	0.902	0.900	0.949	0.941	0.950
		0.30	100	0.952	0.944	0.996	0.993	0.996

$$\eta = \eta_a = 0.5 \text{ and } \eta_e = 1/(1 + \gamma_1)$$

The empirical power is defined as number of times in 20,000 replications the value of the test statistic is greater than the critical value (1.6449). The results are presented in Table 5.

When  $\gamma_1 = 1$  (the null hypothesis), from Theorem 3,  $\eta^* = \eta = 0.5$  ( $\eta^* = \eta_a = \eta_e = 0.5$ ) for the additive (multiplicative) model. Hence, the type I errors of the CA trend tests with  $\eta^*$  and  $\eta$  ( $\eta^*$ ,  $\eta_a$  and  $\eta_e$ ) for the additive (multiplicative) model should be equal. From Table 5, the locally optimal score  $\eta = 0.5$  for the additive model is nearly as powerful as the optimal ones for either the additive or multiplicative models.

## 6. Application

In this section, we apply the results to real data. LIU et al. (2000) conducted a case-control study to examine the association between the variation at the *IL13* gene and atopic dermatitis (AD). The cases consisted of 187 patients from MAS-90 patients with AD and the controls were 98 members of the study population without AD. The data from LIU et al. (2000) is given in Table 6 (*A* is the candidate allele). Penetrance analysis suggested that the genetic model underlying the data in Table 6 is dominant, and the two-sided tests (Pearson  $\chi^2$  and Fisher exact) were applied

Table 6  
Genotype distribution of the *IL13* 4257 G/A polymorphism

	<i>GG</i>	<i>GA</i>	<i>AA</i>	Total
AD	105	72	10	187
Non-AD	68	24	6	98
Total	173	96	16	285

P-values for the CA trend tests

$\eta$	P-values	
	One-sided	Two-sided
0.0	0.6064	0.7872
0.5	0.0478	0.0956
1.0	0.0149	0.0297

yielding  $p$ -values 0.059 and 0.057, respectively (LIU et al., 2000). Applying CA trend tests for several  $\eta$  (0.0, 0.5, 1.0) to the data gave the  $p$ -values presented in Table 6. The results depend on the assumed model with their different scores.

Although the natural alternative hypothesis for CA trend tests with a candidate-gene is one-sided (Appendix A), to compare our analysis with that of LIU et al. (2000), we also consider two-sided CA trend tests. Table 6 shows  $p$ -values of the CA trend test using the optimal score for the dominant model  $\eta = 1.0$  (when  $R \neq S$ ) are more significant than either the Pearson  $\chi^2$  test or Fisher exact test. Indeed, the optimal CA trend test shows that the association between having an A allele and AD is significant.

## 7. Discussion

In this paper, a single parameter is used to describe a family of genetic models and an optimal score for the corresponding CA trend test minimizing the sample size required to achieve a prespecified size and power of the test when  $R = S$  is derived. The locally optimal score test for any  $R$  and  $S$  is also obtained. The results show that the scores of SASIENI (1997) for the recessive and dominant models yield optimal tests while his scores for the additive model corresponds to a locally optimal test. Our parameterization shows that the multiplicative model always lies in between the recessive and additive models and is asymptotically equivalent to the additive model. Examining the performance of locally optimal score ( $\eta = 0.5$ ) for the additive model indicates that it is nearly as powerful as the individual optimal tests. Thus, this commonly used test does not incur a noticeable decrease in power.

## Acknowledgments

The research of Zhaohai Li was supported in part by EY14478 of The National Eye Institute and the work of Joseph Gastwirth was supported in part by SBR-9807731 of The National Science Foundation. We thank two referees for their valuable suggestions, which improved the presentation.

## Appendix A: Proof that $E_{H_a}(U) > 0$

Note  $E_{H_a}(U) = N\mu_a$ . Let  $x = (0, \eta, 1)$  and  $\psi = R/N$ . Then

$$\mu_a = \psi(1 - \psi) \sum_i [x_i(p_i - q_i)] = \psi(1 - \psi)[(p_2 - q_2) - \eta(q_1 - p_1)].$$

From  $0 < f_0 \leq f_1 \leq f_2 < 1$ , it is easy to verify that, under  $H_a$ ,  $f_0 < K < f_2$ . From (5),  $p_2 - q_2 = (f_2 - K)g_2[K(1 - K)]^{-1} > 0$  and  $p_1 - q_1 = (f_1 - K)g_1[K(1 - K)]^{-1}$ . If  $p_1 - q_1 \geq 0$ , then  $\mu_a > 0$ . When  $p_1 - q_1 < 0$ , i.e.,  $f_1 < K$ . Hence

$$\begin{aligned} Kg_0 > f_0g_0 &\Leftrightarrow Kg_0 + f_2g_2 + f_1g_1 > f_0g_0 + f_2g_2 + f_1g_1 = K = Kg_0 + Kg_1 + Kg_2 \\ &\Leftrightarrow f_2g_2 + f_1g_1 > Kg_1 + Kg_2 \Leftrightarrow (f_2 - K)g_2 > (K - f_1)g_1 > 0 \\ &\Leftrightarrow \frac{(f_2 - K)g_2}{(K - f_1)g_1} > 1 \geq \eta. \end{aligned}$$

Since  $q_1 - p_1 > 0$ ,  $(p_2 - q_2)/(q_1 - p_1) > \eta$ , which implies  $(p_2 - q_2) - \eta(q_1 - p_1) > 0$ . Thus,  $\mu_a > 0$ .

## Appendix B: Reparameterization

The family of genetic models defined by  $\theta$  does not automatically include the multiplicative model, which is defined as  $\gamma_2 = \gamma_1^2$  (curve OB in Figure 1). The multiplicative model is locally equivalent to the additive one. Given any  $(\gamma_1, \gamma_2) \in \Omega_\gamma$  such that  $\gamma_2 = \gamma_1^2$ , we have  $\cot \theta = (1 + \gamma_1)^{-1}$ . Hence the multiplicative model is a function of  $\theta$  and  $\gamma_1$ . For the multiplicative model,  $\theta \in [\cot^{-1}(1/(1 + \gamma_1)), \pi/2]$  for any  $\gamma_1 \geq 1$ . This reparameterization shows that the multiplicative model always lies between the additive and recessive models, as the right end point of  $[\cot^{-1}(1/(1 + \gamma_1)), \pi/2]$  corresponds to the recessive model and the left end point corresponds to the additive model when  $\gamma_1 = 1$ . As  $\gamma_1 \rightarrow 1$ , the multiplicative parameters  $(\gamma_1, \gamma_2) \rightarrow (1, 1)$  and  $\cot \theta \rightarrow 1/2$ . Thus, the multiplicative model approaches the additive one and the score for the additive model is locally optimal for a multiplicative model.

## Appendix C: Proof of Theorem 1

We find the score  $x = (0, \eta, 1)$  such that  $\sigma_a^2/\mu_a^2$  is minimized. Let

$$\begin{aligned} M_1(\eta) &= (p_2 - q_2) - \eta(q_1 - p_1) \\ M_2(\eta) &= \eta^2 [(1 - \psi) p_1 - (1 - \psi) p_1^2 + \psi q_1 - \psi q_1^2] \\ &\quad - 2\eta [(1 - \psi) p_1 p_2 + \psi q_1 q_2] \\ &\quad + [(1 - \psi) p_2 - (1 - \psi) p_2^2 + \psi q_2 - \psi q_2^2]. \end{aligned}$$

From (3), it can be shown that  $\mu_a = \psi(1 - \psi) M_1(\eta)$  and  $\sigma_a^2 = \psi(1 - \psi) M_2(\eta)$ . Given  $\psi$ , the problem is to find  $\eta \in [0, 1]$  such that  $N^* = M_2(\eta)/M_1^2(\eta)$  is minimized. By differentiating  $N^*$  with respect to  $\eta$ , we have

$$\frac{\partial}{\partial \eta} N^* = \frac{M_1(\eta) \frac{\partial}{\partial \eta} M_2(\eta) - 2M_2(\eta) \frac{\partial}{\partial \eta} M_1(\eta)}{M_1^3(\eta)},$$

where  $\frac{\partial}{\partial \eta} M_1(\eta) = p_1 - q_1$  and

$$\frac{\partial}{\partial \eta} M_2(\eta) = 2\eta [(1 - \psi) p_1(1 - p_1) + \psi q_1(1 - q_1)] - 2[(1 - \psi) p_1 p_2 + \psi q_1 q_2].$$

After some algebraic manipulation, we obtain

$$\frac{\partial}{\partial \eta} N^* = \frac{2\eta [\psi q_1 + (1 - \psi) p_1] (p_2 q_0 - p_0 q_2) - 2[\psi q_2 + (1 - \psi) p_2] (p_1 q_0 - p_0 q_1)}{M_1^3(\eta)},$$

where  $M_1(\eta) > 0$  (Appendix A) and, from (5),

$$p_2 q_0 - p_0 q_2 = f_0(1 - f_0) g_0 g_2 (\gamma_2 - \delta_2) / [K(1 - K)] > 0$$

and

$$p_1 q_0 - p_0 q_1 = f_0(1 - f_0) g_0 g_1 (\gamma_1 - \delta_1) / [K(1 - K)] \geq 0.$$

Hence, if we set  $\frac{\partial}{\partial \eta} N^* = 0$  and use (5), we obtain a unique solution

$$\eta^* = \left[ \frac{\psi q_2 + (1 - \psi) p_2}{\psi q_1 + (1 - \psi) p_1} \right] \left[ \frac{g_1(\gamma_1 - \delta_1)}{g_2(\gamma_2 - \delta_2)} \right] = \left[ \frac{RK(1 - f_2) + S(1 - K)f_2}{RK(1 - f_1) + S(1 - K)f_1} \right] \cot \theta$$

where  $\gamma_1 - \delta_1 = (\gamma_2 - \delta_2) \cot \theta$ . Since  $\eta^* \geq 0$ , if we show  $\eta^* \leq 1$ , then  $\eta = \eta^*$  minimizes  $N^*$ . To prove  $\eta^* \leq 1$ , recall  $\cot \theta = (f_1 - f_0)/(f_2 - f_0)$ , so

$$\eta^* = \frac{RK(1 - f_2)(f_1 - f_0) + S(1 - K)f_2(f_1 - f_0)}{RK(1 - f_1)(f_2 - f_0) + S(1 - K)f_1(f_2 - f_0)} \leq 1$$

because  $(1 - f_2)(f_1 - f_0) \geq (1 - f_1)(f_2 - f_0)$  and  $f_2(f_1 - f_0) \geq f_1(f_2 - f_0)$ .  $\square$

## Appendix D: Proof of Theorem 2

Case (i) follows from Theorem 1 and  $\cot \theta = 0$  when  $\theta = \pi/2$ , and  $\cot \theta = 1$  and  $f_2 = f_1$  when  $\theta = \pi/4$ . For case (ii), from (6), we need to show  $RK(1 - f_2) + S(1 - K)f_2 = RK(1 - f_1) + S(1 - K)f_1$ , which is equivalent to  $S(f_2 - f_1) = RK(f_2 - f_1) + SK(f_2 - f_1)$ , i.e.,  $S = RK + SK$ , where  $f_2 > f_1$  as  $\cot \theta = (f_1 - f_0)/(f_2 - f_0) < 1$ .  $\square$

## References

- ARMITAGE, P., 1955: Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386.
- COCHRAN, W. G., 1954: Some methods for strengthening the common chi-squared tests. *Biometrics* **10**, 417–451.
- FREIDLIN, B., ZHENG, G., LI, Z. H., and GASTWIRTH, J. L., 2002: Trend tests for case-control studies of genetic markers: Power, sample size and robustness. *Human Heredity* **53**, 146–152.
- GASTWIRTH, J. L. and FREIDLIN, B., 2000: On power and efficiency robust linkage tests for affected sibs. *Annals of Human Genetics* **64**, 443–453.
- GRAUBARD, B. I. and KORN, E. L., 1987: Choice of column scores for testing independence in ordered  $2 \times K$  contingency tables. *Biometrics* **43**, 471–476.
- LIU, X., NICKEL, R., BEYER, K., WAHN, U., EHRLICH, E., FREIDHOFF, L. R., BJORKSTEN, B., BEATY, T., and HUANG, S.-K., 2000: An IL-13 coding region variant is associated with a high total serum IgE level and atopic dermatitis in the German Multicenter Atopy Study (MAS-90). *Journal of Allergy and Clinical Immunology* **106**, 167–170.
- NEUHAUSER, M. and HOTHORN, L. A., 1999: An exact Cochran-Armitage test for trend when dose-response shapes are a priori unknown. *Computational Statistics and Data Analysis* **30**, 403–412.
- PODGOR, M. J., GASTWIRTH J. L., and MEHTA C. R., 1996: Efficiency robust tests of independence in contingency tables with ordered classifications. *Statistics in Medicine* **15**, 2095–2105.
- SASIENI, P. D., 1997: From genotypes to genes: Doubling the sample size. *Biometrics* **53**, 1253–1261.
- SHIH, M. and WHITTEMORE, A. S., 2001: Allele-sharing among affected relatives: non-parametric methods for identifying genes. *Statistical Methods in Medical Research* **10**, 27–55.
- SLAGER, S. L. and SCHAID, D. J., 2001: Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Human Heredity* **52**, 149–153.
- TARONE, R. E. and GART, J. J., 1980: On the robustness of combined tests for trends in proportions. *Journal of the American Statistical Association* **75**, 110–116.
- WHITTEMORE, A. S. and TU, I.-P., 1998: Simple, robust linkage tests for affected sibs. *American Journal of Human Genetics* **62**, 1228–1242.

Received April 2002  
 Revised May 2002  
 Revised October 2002  
 Accepted November 2002